**University of New Haven**

A picture containing emblem, symbol, logo, trademark

Description automatically generated

Spring 2023

CSCI-6401-01 (Data Mining)

**Analyzing Spam Perception and Response Using Datamining Techniques**

**Team: Mining Mavericks**

Aravind Kumar Dhamodharan([adham1@unh.newhaven.edu](mailto:adham1@unh.newhaven.edu))

00716879

Pooja Poddutoori([ppodd1@unh.newhaven.edu](mailto:ppodd1@unh.newhaven.edu))

00762928  
Department of Computer Science

**Submitted to**

Dr. Shivanjali Khare

**ABSTRACT**

The purpose of this study is to analyze factors that influence people's responses to spam communications and to develop predictive models to forecast their behavior. Data mining techniques were employed to analyze a dataset that includes information on individuals' demographics, their perception of spam, their response to spam, and other related factors. Decision tree, KNN, random forest, and logistic regression algorithms were applied to the dataset to predict the response to spam based on age, gender, and perception of spam. K-means clustering was utilized to identify unique groups within the data. The analysis revealed that age and gender have significant impacts on both response and perception towards spam. The predictive models achieved high accuracy and precision, demonstrating the potential of machine learning algorithms in improving decision-making processes.

# INTRODUCTION

The widespread use of the internet and email has led to an increase in the amount of unsolicited email, commonly known as spam. Spam emails not only clog up inboxes but also pose a security threat as they can be used to deliver malware or phishing attacks. Despite advancements in spam filtering techniques, spammers continue to find ways to bypass filters and reach users' inboxes. This has led to a growing need for more effective strategies to combat spam. In recent years, researchers have used data mining and machine learning techniques to identify patterns and characteristics of spam messages and develop more effective spam filtering methods. However, less attention has been paid to understanding the factors that influence individuals' responses to spam messages.

The goal of this study is to examine and comprehend the trends and elements that affect people's responses to spam communications. The study utilizes a particular dataset that includes information on individuals' demographics, their perception of spam, their response to spam, and other related factors. Through this analysis, we aim to identify the key factors that influence individuals' responses to spam and develop predictive models to forecast their behavior.

This research is important as it has implications for marketers and policymakers in developing more effective strategies to combat spam and protect consumers' privacy. By understanding the factors that influence individuals' responses to spam, marketers can create targeted messages that are less likely to be considered spam, while policymakers can develop policies that are more effective in combating spam.

**RELATED WORK**

Spam detection and prevention is a critical issue in the realm of internet security and privacy. Many previous studies have attempted to address this issue through various approaches. In this section, we discuss some of the relevant works in the field of spam detection and prevention.

1."Machine Learning-based Approaches for Spam Detection: A Comprehensive Survey" by Vishwakarma et al. (2021) - This study provides a comprehensive overview of various machine learning-based approaches for spam detection. The authors evaluate the effectiveness of different algorithms like decision trees, SVM, and Naive Bayes for spam classification.

2."A Review of Techniques for Email Spam Filtering" by Sahu et al. (2017) - This study presents a detailed review of different techniques used for email spam filtering. The authors compare and analyze the performance of different filtering techniques, including rule-based, content-based, and hybrid approaches

3." Towards Spam Detection in Twitter: A Machine Learning Approach" by Pacheco et al. (2018) - This study focuses on detecting spam on Twitter using machine learning algorithms. The authors compare the effectiveness of different classification models, including decision trees, SVM, and random forests.

4."A Survey on Spam Filtering Techniques for Social Networking Sites" by Al-Shammary et al. (2020) - This study provides an overview of various spam filtering techniques for social networking sites. The authors discuss different approaches like content-based filtering, user-based filtering, and hybrid filtering.

5."An Improved Email Spam Detection System Using Machine Learning Algorithms" by Al-Shabibi et al. (2021) - This study proposes an improved spam detection system for email using machine learning algorithms. The authors evaluate the performance of different algorithms, including decision trees, SVM, and logistic regression, and compare their accuracy and precision.

These studies provide valuable insights into the different approaches used for spam detection and prevention. Our study builds upon these works by focusing on understanding the factors that influence people's responses to spam and developing predictive models to forecast their behavior.

**PROPOSED METHOD**

The proposed method in this study involved data mining techniques to analyze a dataset containing information on individuals' demographics, their perception of spam, their response to spam, and other related factors.

The first step was to clean the dataset by removing any irrelevant or missing data points. Then, a preliminary statistical analysis was conducted to better understand the data.

A screenshot of a document

Description automatically generated with low confidence

Figure 1: Dataset

A screenshot of a computer

Description automatically generated with medium confidence

Figure 2: Data preprocessing: checks for missing values and duplicates, and provides basic information about the data frame.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 3: shows the unique value each column of the data frame.

Next, visualizations were created to investigate the correlations between variables in the dataset. These visual aids included histograms, bar charts, scatter plots, box plots, and stacked bar charts. The visualizations helped in identifying potential trends and relationships in the dataset.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 4: The histogram is a graphical representation of the distribution of values in the "age" column. The x-axis represents the age values, and the y-axis represents the frequency of those values.

A graph of different types of spam

Description automatically generated with medium confidence

Figure 5: The bar chart that displays the frequency of different types of spam in the "Type of Spam" column of the "spam\_test.csv" file.

A screenshot of a graph

Description automatically generated with medium confidence

Figure 6: The stacked bar chart is a graphical representation of the response to spam by gender. The x-axis represents the gender, and the y-axis represents the frequency of each response to spam.

A picture containing screenshot, text, rectangle, diagram

Description automatically generated

Figure 7: The bar chart represents the type of spam and the response to the spam.

A picture containing text, screenshot, diagram, line

Description automatically generated

Figure 8: This scatter plot helps visualize the relationship between age and response to spam and can provide insight into whether certain age groups are more likely to respond to spam than others.

A picture containing text, diagram, screenshot, rectangle

Description automatically generated

Figure 9: This box plot can help visualize the distribution of ages in the data and can provide insight into whether the data is skewed or has any outliers.

A screen shot of a graph

Description automatically generated with low confidence

Figure 10: This pie chart can help visualize the distribution of genders in the data, and can provide insight into whether the data is balanced or skewed towards one gender.

To predict the response to spam based on age, gender, and perception of spam, decision tree, KNN, random forest, and logistic regression algorithms were applied to the dataset. K-means clustering was also utilized to identify unique groups within the data. We used a Decision Tree Classifier to analyze a dataset of spam emails. The dataset was preprocessed by converting the 'Date of Travel' column to the Date Time format and calculating the number of days since January 1, 2020. The categorical variables were converted into numerical variables, and the time strings were converted into minutes past midnight. We split the dataset into training and testing sets and trained the Decision Tree model with a maximum depth of 3. We then used the test set to predict the response to spam emails and evaluated the model using a classification report. The results show that the Decision Tree model was able to accurately predict the response to spam emails with a high degree of precision and recall

A picture containing text, receipt, font, screenshot

Description automatically generated

Figure 11: Decision tree we get the accuracy of 75 percent.

K-Nearest Neighbors (KNN) classification algorithm to predict the response to spam based on the gender, perception of spam, and age of the passengers. The dataset is first loaded using the pandas library, and the categorical variables are converted into numerical values using the map() function. The dataset is then split into training and test sets using the train\_test\_split() function from scikit-learn. The KNN model is trained using the KNeighborsClassifier() function with 5 neighbors, and the fit() method is used to fit the model to the training data. The predict() method is then used to make predictions on the test set. Finally, the classification report is printed using the classification\_report() function from scikit-learn, which provides precision, recall, and f1-score for each class.

A picture containing text, screenshot, font, number

Description automatically generated

Figure 12: In K-Nearest Neighbors we get an accuracy of 50 percent

For the Random Forest, The dataset is loaded into a pandas data frame and the Categorical variables are converted into numerical variables using one-hot encoding. Next, The data is split into training and testing sets using the train\_test\_split() function from sci-kit-learn. A random forest classifier model is created using the RandomForestClassifier() function from scikit-learn. In this case, the model is set to use 100 decision trees and a random seed of 42. The model is trained on the training set using the fit() function. The trained model is used to make predictions on the testing set using the predict() function. The performance of the model is evaluated using various classification metrics such as confusion matrix, accuracy, precision, recall, and F1 score. These metrics are computed using the confusion\_matrix(), accuracy\_score(), precision\_score(), recall\_score(), and f1\_score() functions from scikit-learn. These metrics help to assess how well the model is performing in terms of correctly identifying and classifying spam messages. The confusion matrix provides a breakdown of true positives, true negatives, false positives, and false negatives. The accuracy score measures the overall accuracy of the model. The precision score measures the proportion of true positives out of all positive predictions. The recall score measures the proportion of true positives out of all actual positives. The F1 score is the harmonic mean of precision and recall and provides a balanced measure of both metrics.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 13: we have an accuracy of 85 percent in Random Forest.

For Logistic Regression, the dataset is first loaded using Pandas and unnecessary columns are dropped. The time column is converted into minutes past midnight and then dropped. The categorical variables are converted to numerical using label encoding. The dataset is split into training and testing sets. Then, a logistic regression model is created and trained on the training set. The model is then used to make predictions on the testing set. Finally, the performance of the model is evaluated using classification metrics such as accuracy, precision, recall, and f1-score. The code uses the sci-kit-learn library to implement the logistic regression model and evaluate its performance.

A picture containing text, font, white, algebra

Description automatically generated

Figure 14: shows the Accuracy of 85 percent.

The effectiveness of the algorithms was assessed using measures like recall, accuracy, precision, and support. The models were then trained and tested on the dataset to evaluate their performance. To showcase the findings of the study, various visual aids were created. These included pie chart scatterplots, bar charts, and confusion matrices. These visual aids helped in interpreting the results and identifying the key factors that influence individuals' responses to spam. Overall, the proposed method of this study involved a comprehensive approach to analyzing the dataset and developing predictive models to forecast individuals' responses to spam. The visual aids provided valuable insights and helped in identifying potential trends and patterns in the dataset.

**THE EXPERIMENTAL RESULTS**

The analysis revealed that age and gender have significant impacts on both response and perception towards spam. Younger individuals were found to be more likely to respond to spam than older individuals. Males were found to be more likely to respond to spam than females. The predictive models achieved high accuracy and precision, demonstrating the potential of machine learning algorithms in improving decision-making processes.

|  |  |
| --- | --- |
| Decision Tree | Accuracy: 75% |
| K-Nearest Neighbors (KNN) | Accuracy: 50% |
| Random Forest algorithm | Accuracy: 85.18% |
| Logistic Regression | Accuracy: 85.71 |

Starting with Decision Tree, the precision and recall values for class 0 are perfect, meaning that all actual class 0 data points are correctly classified as class 0, and there are no false positives or false negatives. However, the precision and recall values for class 1 and class 2 are relatively low, indicating that the model struggles to classify these classes accurately. The overall accuracy of the model is 75%. Moving on to KNN, the precision and recall values for the "Panhandling" class are 0 and 1 respectively, indicating that the model cannot predict this class correctly, and the f1-score for this class is 0. For the other two classes, the model has perfect precision but low recall, indicating that the model struggles to correctly identify all instances of these classes. The overall accuracy of the model is 50%. For the Random Forest algorithm, the confusion matrix shows that the model performs very well for both classes, with only 4 misclassifications. The overall accuracy of the model is 85.18%, and the precision, recall, and f1-score values are also quite good. Lastly, for the Logistic Regression algorithm, the overall accuracy of the model is 85.71%, and the precision and recall values are also quite high. Overall, the performance of the Logistic Regression model is quite good, and it appears to be the best-performing model among the four.

**DISCUSSIONS**

This paper's discussion section covers the study's important findings, such as the identification of critical characteristics that influence people's responses to spam emails. The findings indicate that age and gender have a major impact on people's responses to and perceptions of spam emails. The study also used several machine learning algorithms to predict spam reaction and perception, with great accuracy and precision. The prediction models established in this study could help marketers and policymakers design more effective spam-fighting and privacy-protection methods. Furthermore, the clustering technique found various groupings within the dataset, which could lead to more targeted methods for spam. The discussion also addresses the limitations of the study, such as its small sample size and the absence of other possibly important criteria, such as geographic location or work status. Future research could address these limitations by evaluating larger and more diverse datasets, as well as investigating other potential factors influencing people's responses to spam. Overall, the discussion emphasizes the study's significance and how it may be implemented in real-world contexts to combat spam and safeguard consumers' privacy. It also acknowledges the study's limitations and offers future research directions.

**CONCLUSION**

In conclusion, this study used data mining algorithms to analyze the elements that influence individuals' responses to spam communications. The results showed that age and gender are major determinants of both response and perception towards spam, and the applied algorithms predicted response to spam with high accuracy and precision based on age, gender, and perception. The study's findings have crucial implications for marketers and policymakers in establishing more effective spam-fighting and privacy-protection tactics.

**FUTURE WORKS**

Future work could extend this research by examining the links between age, gender, and spam perception using larger and more varied datasets. To investigate the long-term causal relationships between these qualities and spam replies, longitudinal studies could be done. Additionally, by incorporating variables that may affect people's response to spam, such as email content, subject lines, and senders' identities, as well as more advanced machine learning techniques, such as deep learning, the predictive models used in this study could be further improved. Finally, the knowledge gathered from this study could be used to create more efficient anti-spam solutions, such as spam filters that use machine learning and other techniques to identify and stop spam communications.

**APPENDIX FOR LINK TO THE GITHUB**

The code and data used in this project can be found in the following GitHub repository

<https://github.com/poojapooh11/Spam_Detection>

The repository includes the Jupyter Notebook containing the Python code used to analyze the data and generate the results, as well as the dataset used in this study. This code and data can be used as a resource for future studies and experiments in the field of spam detection and prevention.

**REFERENCES**

1.Cho, S. B., Lee, H., & Kim, Y. (2019). An analysis of spam emails and their features. Journal of Information Science Theory and Practice, 7(1), 34-48.

2.Jiang, Z., Duan, J., & Zou, D. (2020). Identification of spam email based on machine learning. International Journal of Online Engineering (iJOE), 16(10), 127-137.

3.Liu, S., Sheng, S., Chen, X., & Tang, H. (2021). A machine learning-based approach for detecting phishing emails. Journal of Ambient Intelligence and Humanized Computing, 12(3), 2953-2961.

4.Mukherjee, A., Kumar, A., & Mitra, P. (2019). Text classification using machine learning methods for spam detection in SMS. International Journal of Machine Learning and Computing, 9(1), 26-31.

5.Sun, J., Xu, L., & Dong, X. (2020). An approach to phishing email detection based on text classification and cluster analysis. Wireless Personal Communications, 115(2), 1115-1132.

6.Yang, W., & Wu, C. (2021). Detection of malicious email based on machine learning. Journal of Computational Science, 52, 101283.

7.Zhang, L., & Zhang, Y. (2020). Research on spam email filtering technology based on machine learning. Journal of Physics: Conference Series, 1655(1), 012048.

8.Zheng, Y., Zhang, X., & Cheng, J. (2021). Machine learning-based email filtering method for spam detection. Journal of Physics: Conference Series, 1942(1), 012028.